

Beyond the Hazard Rate: More Perturbation Algorithms for Adversarial Multi-armed Bandits

Zifan Li
University of Michigan
zifanli@umich.edu

Ambuj Tewari
University of Michigan
tewaria@umich.edu

February 21, 2017

Abstract

Recent work on follow the perturbed leader (FTPL) algorithms for the adversarial multi-armed bandit problem has highlighted the role of the hazard rate of the distribution generating the perturbations. Assuming that the hazard rate is bounded, it is possible to provide regret analyses for a variety of FTPL algorithms for the multi-armed bandit problem. This paper pushes the inquiry into regret bounds for FTPL algorithms beyond the bounded hazard rate condition. There are good reasons to do so: natural distributions such as the uniform and Gaussian violate the condition. We give regret bounds for both bounded support and unbounded support distributions without assuming the hazard rate condition. We also disprove a conjecture that the Gaussian distribution cannot lead to a low-regret algorithm. In fact, it turns out that it leads to near optimal regret, up to logarithmic factors. A key ingredient in our approach is the introduction of a new notion called the generalized hazard rate.

1 Introduction

Starting from the seminal work of [Hannan \[1957\]](#) and later developments due to [Kalai and Vempala \[2005\]](#), perturbation based algorithms (called “Follow the Perturbed Leader (FTPL)”) have occupied a central place in online learning. Another major family of online learning algorithms, called “Follow the Regularized Leader (FTRL)”, is based on the idea of regularization. In special cases, such as the exponential weights algorithm for the experts problem, it has been folk knowledge that regularization and perturbation ideas are connected. That is, the exponential weights algorithm can be understood as either using negative entropy regularization or Gumbel distributed perturbation (for example, see the discussion in [Abernethy et al. \[2014\]](#)).

Recent work have begun to further uncover the connections between perturbation and regularization. For example, in online linear optimization, one can understand regularization and perturbation as simply two different ways to smooth a non-smooth potential function. The former corresponds to infimal convolution smoothing and the latter corresponds to stochastic (or integral convolution) smoothing [[Abernethy et al., 2014](#)]. Having a generic framework for understanding perturbations allows one to study a wide variety of online linear optimization games and a number of interesting perturbations.

FTRL and FTPL algorithms have also been used beyond “full information” settings. “Full information” refers to the fact that the learner observes the entire move of the adversary. The multi-armed bandit problem is one of the most fundamental examples of “partial information” settings. Regret analysis of the multi-armed bandit problem goes back to the work of [Robbins \[1952\]](#) who formulated the stochastic version of the problem. The non-stochastic, or adversarial, version was formulated by [Auer et al. \[2002\]](#), who provided the EXP3 algorithm achieving $O(\sqrt{NT \log N})$ regret in T rounds with N arms. They also showed a lower bound of $\Omega(\sqrt{NT})$, which was later matched by the Poly-INF algorithm [[Audibert and Bubeck, 2009](#), [Audibert et al., 2011](#)]. The Poly-INF algorithm can be interpreted as an FTRL algorithm with negative Tsallis entropy regularization [[Audibert et al., 2011](#), [Abernethy et al., 2015](#)]. For a recent survey of both stochastic and non-stochastic bandit problems, see [Bubeck and Cesa-Bianchi \[2012\]](#).

For the non-stochastic multi-armed bandit problem, [Kujala and Elomaa \[2005\]](#) and [Poland \[2005\]](#) both showed that using the exponential (actually double exponential/Laplace) distribution in an FTPL algorithm coupled with standard unbiased estimation technique yields near-optimal $O(\sqrt{NT \log N})$ regret. Unbiased estimation needs access to arm probabilities that are not explicitly available when using an FTPL algorithm. [Neu and Bartók \[2013\]](#) introduced the geometric resampling scheme to approximate these probabilities while still guaranteeing low regret. Recently, [Abernethy et al. \[2015\]](#) analyzed FTPL for adversarial multi-armed bandits and provided regret bounds under the condition that the hazard rate of the perturbation distribution is bounded. This condition allowed them to consider a variety of perturbation distributions beyond the exponential, such as Gamma, Gumbel, Frechet, Pareto, and Weibull.

Unfortunately, the bounded hazard rate condition is violated by two of the most widely known distributions: namely the uniform¹ and the Gaussian distributions. As a result, the results of [Abernethy et al. \[2015\]](#) say nothing about the regret incurred in an adversarial multi-armed bandit problem when we use these distributions to generate perturbations. Contrast this to the full information experts setting where using these distributions as perturbations yields optimal \sqrt{T} regret and even yields the optimal $\sqrt{\log N}$ dependence on the dimension in the Gaussian case [[Abernethy et al., 2014](#)].

The Gaussian distribution has lighter tails than the exponential. The hazard rate of a Gaussian increases linearly on the real line (and is hence unbounded) whereas the exponential has a constant hazard rate. Does having too light a tail makes a perturbation inherently bad? The uniform is even worse from a light tail point of view: it has bounded support! In fact, [Kujala and Elomaa \[2005\]](#) had trouble dealing with the uniform distribution and remarked, “we failed to analyze the expert setting when the perturbation distribution was uniform.” Does having a bounded support make a perturbation even worse? Or is it that the hazard rate condition is just a sufficient condition without being anywhere close to necessary for a good regret bound to exist. The analysis of [Abernethy et al. \[2015\]](#) suggests that perhaps a bounded hazard rate is critical. They even made the following conjecture.

Conjecture 1. *If a distribution \mathcal{D} has a monotonically increasing hazard rate $h_{\mathcal{D}}(x)$ that does not converge as $x \rightarrow +\infty$ (e.g., Gaussian), then there is a sequence of losses that incur at least a linear regret.*

The main contribution of this paper is to provide answers to the questions raised above. First, we show that boundedness of the hazard rate is certainly not a requirement for achieving sublinear (in T) regret. Bounded support distributions, like the uniform, violate the boundedness condition on the hazard rate in the most extreme way. Their hazard rate blows up not just asymptotically at infinity, as in the Gaussian case, but as one approaches the right edge of the support. Yet, we can show (Corollary 3.3) that using the uniform distribution results in a regret bound of $O((NT)^{2/3})$. This bound is clearly not optimal. But optimality is not the point here. What is surprising, especially if one regards Conjecture 1 as plausible, is that a non-trivial sublinear bound holds at all. In fact, we show (Corollary 3.4) that using *any* continuous distribution with bounded support and bounded density results in a sublinear regret bound.

Second, moving beyond bounded support distributions to ones with unbounded support, we settle Conjecture 1 in the negative. In Theorem 4.6 we show that, instead of suffering linear regret as predicted by Conjecture 1, a perturbation algorithm using the Gaussian distribution enjoys a near optimal regret bound of $O(\sqrt{NT \log N \log T})$. A key ingredient in our approach is a new quantity that we call the *generalized hazard rate* of a distribution. We show that bounded generalized hazard rate is enough to guarantee sublinear regret in T (Theorem 4.2).

Finally, we investigate the relationship between tail behavior of random perturbations and the regret they induce. We show that heavy tails, along with some fairly mild assumptions, guarantee a bounded hazard rate (Theorem 4.9) and hence previous results can yield regret bounds for these perturbations. However, light tails can fail to have a bounded hazard rate. Nevertheless, we show that under reasonable conditions, light tailed distributions do have a bounded *generalized* hazard rate (Theorem 4.10). This result allows us to show that reasonably behaved light-tailed distributions lead to near optimal regret (Corollary 4.11). In particular, the exponential power (or generalized normal) family of distributions yields near optimal regret (Theorem 4.13).

Due to space restrictions, all proofs are deferred to the appendix.

¹The uniform distribution is also historically significant as it was used in the original FTPL algorithm of [Hannan \[1957\]](#).

2 Follow the Perturbed Leader Algorithm for Bandits

Recall the setting of the adversarial multi-armed bandit problem [Auer et al., 2002]. An adversary (or Nature) chooses loss vectors $g_t \in [-1, 0]^N$ for $1 \leq t \leq T$ ahead of the game. Such an adversary is called *oblivious*. At round $t = 1, \dots, T$ in a repeated game, the learner must choose a distribution $p_t \in \Delta_N$ over the set of N available arms (or actions). The learner plays action i_t sampled according to p_t and incurs the loss $g_{t,i_t} \in [-1, 0]$. The learner observes only g_{t,i_t} and receives no information about the values $g_{t,j}$ for $j \neq i_t$.

The learner's goal is to minimize the *regret*. Regret is defined to be the difference in the realized loss and the loss of the best fixed action in hindsight:

$$\text{Regret}_T := \max_{i \in [N]} \sum_{t=1}^T (g_{t,i} - g_{t,i_t}). \quad (1)$$

To be precise, we consider the *expected* regret, where the expectation is taken with respect to the learner's randomization. Note that, under an oblivious adversary, the only random variables in the above expression are the actions i_t of the learner.

The maximization in (1) implies that g is strictly speaking a negative *gain* vector, not a loss vector. Nevertheless, we use the term *loss*, as we impose the assumption that $g_t \in [-1, 0]^N$ throughout the paper. The decision to consider the loss setting is important: our proof will not work for gains. It is known that the adversarial multi-armed bandit problem does not exhibit symmetry with respect to gains versus losses. Often losses are easier to handle than gains [Bubeck and Cesa-Bianchi, 2012]. Finally, our decision to treat losses as negative gains stems from the desire to work with convex, not concave, potential functions.

2.1 The Gradient-Based Algorithmic Template

We will consider the algorithmic template described in Framework 1, which is the Gradient Based Prediction Algorithm (GBPA) (see, for example, Abernethy et al. [2015]). Let Δ^N be the $(N - 1)$ -dimensional probability simplex in \mathbb{R}^N . Denote the standard basis vector along the i th dimension by e_i . At any round t , the action choice i_t is made by sampling from the distribution p_t which is obtained by applying the gradient of a convex function $\tilde{\Phi}$ to the estimate \hat{G}_{t-1} of the cumulative gain vector so far. The choice of $\tilde{\Phi}$ is flexible but it must be a differentiable convex function such that its gradient is always in Δ^N .

Note that we do not require that the range of $\nabla \tilde{\Phi}$ be contained in the *interior* of the probability simplex. If we required the gradient to lie in the interior, we would not be able to deal with bounded support distributions such as the uniform distribution. Even though some entries of the probability vector p_t might be 0, the estimation step is always well defined since $p_{t,i_t} > 0$. But allowing $p_{t,i}$ to be zero means that \hat{g}_t is not exactly an unbiased estimator of g_t . Instead, it is an unbiased estimator on the support of p_t . That is, $\mathbb{E}[\hat{g}_{t,i} | i_{1:t-1}] = g_{t,i}$ for any i such that $p_{t,i} > 0$. Here, $i_{1:t-1}$ is shorthand for i_1, \dots, i_{t-1} . Therefore, irrespective of whether $p_{t,i} = 0$ or not, we always have

$$\mathbb{E}[p_{t,i} \hat{g}_{t,i} | i_{1:t-1}] = p_{t,i} g_{t,i}. \quad (2)$$

When $p_{t,i} = 0$, we have $\hat{g}_{t,i} = 0$ but $g_{t,i} \leq 0$, which means that \hat{g}_t overestimates g_t outside the support of p_t . Hence, we also have

$$\mathbb{E}[\hat{g}_t | i_{1:t-1}] \succeq g_t, \quad (3)$$

where \succeq means element-wise greater than.

We now present a basic result bounding the expected regret of GBPA in the multi-armed bandit setting. It is basically just a simple modification of the arguments in Abernethy et al. [2015] to deal with the possibility that $p_{t,i} = 0$. We state and prove this result here for completeness without making any claim of novelty.

Framework 1: Gradient-Based Prediction Alg. (GBPA) Template for Multi-Armed Bandits.

GBPA($\tilde{\Phi}$): $\tilde{\Phi}$ is a differentiable convex function such that $\nabla\tilde{\Phi} \in \Delta^N$

Nature: Adversary chooses “gain” vectors $g_t \in [-1, 0]^N$ for $t = 1, \dots, T$

Learner initializes $\hat{G}_0 = 0$

for $t = 1$ to T **do**

Sampling: Learner chooses i_t according to the distribution $p_t = \nabla\tilde{\Phi}(\hat{G}_{t-1})$

Cost: Learner incurs (and observes) “gain” $g_{t,i_t} \in [-1, 0]$

Estimation: Learner creates estimate of gain vector $\hat{g}_t := \frac{g_{t,i_t}}{p_{t,i_t}} \mathbf{e}_{i_t}$

Update: Cumulative gain estimate so far $\hat{G}_t = \hat{G}_{t-1} + \hat{g}_t$

end for

Lemma 2.1. (Decomposition of the Expected Regret) Define the non-smooth potential $\Phi(G) = \max_i G_i$. The expected regret of GBPA($\tilde{\Phi}$) can be written as

$$\mathbb{E}\text{Regret}_T = \Phi(G_T) - \mathbb{E} \left[\sum_{t=1}^T \langle p_t, g_t \rangle \right]. \quad (4)$$

Furthermore, the expected regret of GBPA($\tilde{\Phi}$) can be bounded by the sum of an overestimation, an underestimation, and a divergence penalty:

$$\mathbb{E}\text{Regret}_T \leq \underbrace{\tilde{\Phi}(0)}_{\text{overestimation penalty}} + \mathbb{E} \left[\underbrace{\Phi(\hat{G}_T) - \tilde{\Phi}(\hat{G}_T)}_{\text{underestimation penalty}} \right] + \mathbb{E} \left[\sum_{t=1}^T \underbrace{\mathbb{E}[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) | i_{1:t-1}]}_{\text{divergence penalty}} \right], \quad (5)$$

where the expectations are over the sampling of i_t and $D_{\tilde{\Phi}}$ is the Bregman divergence induced by $\tilde{\Phi}$.

2.2 Stochastic Smoothing of Potential Function

Let \mathcal{D} be a continuous distribution with finite expectation, probability density function f , and cumulative distribution function F . Consider GBPA with potential function of the form:

$$\tilde{\Phi}(G; \mathcal{D}) = \mathbb{E}_{Z_1, \dots, Z_N \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}} \Phi(G_i + Z_i), \quad (6)$$

which is a *stochastic smoothing* of the non-smooth function $\Phi(G) = \max_i G_i$. We will often hide the dependence on the distribution \mathcal{D} if the distribution is obvious from the context or when the dependence on \mathcal{D} is not of importance in the argument. Since Φ is convex, $\tilde{\Phi}$ is also convex. For stochastic smoothing, we have the following result to control the underestimation and overestimation penalty.

Lemma 2.2. For any G , we have

$$\Phi(G) + \mathbb{E}[Z_1] \leq \tilde{\Phi}(G) \leq \Phi(G) + \text{EMAX}(N) \quad (7)$$

where $\text{EMAX}(N)$ is any function such that

$$\mathbb{E}_{Z_1, \dots, Z_N} [\max_i Z_i] \leq \text{EMAX}(N).$$

In particular, this implies that the overestimation penalty $\tilde{\Phi}(0)$ is upper bounded by $\Phi(0) + \text{EMAX}(N) = \text{EMAX}(N)$ and the underestimation penalty $\Phi(\hat{G}_T) - \tilde{\Phi}(\hat{G}_T)$ is upper bounded by $-\mathbb{E}[Z_1]$.

Note that Φ is differentiable with probability 1 (under the randomness of the Z_i 's) due to the fact that Z_i 's are random variables with a density. By Proposition 2.3 of Bertsekas [1973], we can swap the order of differentiation and expectation:

$$\nabla \tilde{\Phi}(G; \mathcal{D}) = \mathbb{E}_{Z_1, \dots, Z_N \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}} e_{i^*}, \text{ where } i^* = \arg \max_{i=1, \dots, N} \{G_i + Z_i\}. \quad (8)$$

Note that, for any G , the random index i^* is unique with probability 1. Hence, ties between arms can be resolved arbitrarily. It is clear from above that $\nabla\tilde{\Phi}$, being an expectation of vectors in the probability simplex, is in the probability simplex. Thus, it is a valid potential to be used in Framework 1. Note that

$$\begin{aligned}\nabla_i\tilde{\Phi}(G) &= \frac{\partial\tilde{\Phi}}{\partial G_i} = \mathbb{E}_{Z_1,\dots,Z_N}\mathbf{1}\{G_i + Z_i > G_j + Z_j, \forall j \neq i\} \\ &= \mathbb{E}_{\tilde{G}_{-i}}[\mathbb{P}_{Z_i}[Z_i > \tilde{G}_{-i} - G_i]] = \mathbb{E}_{\tilde{G}_{-i}}[1 - F(\tilde{G}_{-i} - G_i)].\end{aligned}\tag{9}$$

where $\tilde{G}_{-i} = \max_{j \neq i} G_j + Z_j$. If \mathcal{D} has unbounded support then this partial derivative is non-zero for all i given any G . However, it can be zero if \mathcal{D} has bounded support. Moreover, we have the following useful identity that writes the Hessian of the smoothed potential function in terms of the expectation of the probability density function.

$$\begin{aligned}\nabla_{ii}^2\tilde{\Phi}(G) &= \frac{\partial}{\partial G_i}\nabla_i\tilde{\Phi}(G) = \frac{\partial}{\partial G_i}\mathbb{E}_{\tilde{G}_{-i}}[1 - F(\tilde{G}_{-i} - G_i)] \\ &= \mathbb{E}_{\tilde{G}_{-i}}\left[\frac{\partial}{\partial G_i}(1 - F(\tilde{G}_{-i} - G_i))\right] = \mathbb{E}_{\tilde{G}_{-i}}f(\tilde{G}_{-i} - G_i).\end{aligned}\tag{10}$$

2.3 Connection to Follow the Perturbed Leader

The sampling step of Framework 1 with a stochastically smoothed Φ as the potential $\tilde{\Phi}$ (Equation 6) can be done efficiently. Instead of evaluating the expectation (Equation 8), we just take a random sample. Doing so gives us an equivalent of Follow the Perturbed Leader Algorithm (FTPL) [Kalai and Vempala, 2005] applied to the bandit setting. On the other hand, the estimation step is hard because generally there is no closed-form expression for $\nabla\tilde{\Phi}$.

To address this issue, Neu and Bartók [2013] proposed Geometric Resampling (GR), an iterative resampling process to estimate $\nabla\tilde{\Phi}$ (with bias). They showed that the extra regret after stopping at M iterations of GR introduces an estimation bias that is at most $\frac{NT}{eM}$ as an additive term. That is, all GBPA regret bounds that we prove will hold for the corresponding FTPL algorithm that does M iterations of GR at every time step, with an extra additive $\frac{NT}{eM}$ term. This extra term does not affect the regret rate as long as $M = \sqrt{NT}$, because the lower bound for any adversarial multi-armed bandit algorithm is of the order \sqrt{NT} .

2.4 The Role of the Hazard Rate and Its limitation

In previous work, Abernethy et al. [2015] proved that for a continuous random variable Z with finite and nonnegative expectation and support on the whole real line \mathbb{R} , if the hazard rate of the random variable is bounded, i.e.,

$$\sup_z \frac{f(z)}{1 - F(z)} < \infty,$$

then the expected regret of GBPA can be upper bounded as

$$\mathbb{E}\text{Regret}_T = O\left(\sqrt{NT \times \text{EMAX}(N)}\right).$$

Common families of distributions whose regret can be controlled in this way include Gumbel, Frechet, Weibull, Pareto, and gamma (see Abernethy et al. [2015] for details). However, there are many other families of distributions where the hazard rate condition fails. For example, if the random variable has a bounded support, then the hazard rate would certainly explode at the end of the support. This is, in some sense, an extreme case of violation because the random variable does not even have a tail. There are also some random variables that do have support on \mathbb{R} but have unbounded hazard rate, e.g. Gaussian, where the hazard rate monotonically increases to infinity. How can we perform analyses of the expected regret of GBPA using those random variables as perturbations? To address these issues, we need to go beyond the hazard rate.

3 Perturbations with Bounded Support

In this section, we prove that GBPA with any continuous distribution that has bounded support, bounded density enjoys sublinear expected regret. From Lemma 2.1 we see that the expected regret can be upper bounded by the sum of three terms. The overestimation penalty can be bounded very easily via Lemma 2.2 for a distribution with bounded support. The underestimation penalty is non-positive as long as the distribution has non-negative expectation. The only term that needs to be controlled with some effort is the divergence penalty.

We first present a general lemma that allows us to write the divergence penalty under a stochastic smoothing potential $\tilde{\Phi}$ as a sum involving certain double integrals.

Lemma 3.1. *When using a stochastically smoothed potential as in (6), the divergence penalty can be written as*

$$\mathbb{E} \left[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) | i_{1:t-1} \right] = \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{\left| \frac{g_{t,i}}{p_{t,i}} \right|} \mathbb{E}_{\hat{G}_{-i}} \left[\int_0^s f(\hat{G}_{-i} - \hat{G}_{t-1,i} + r) dr \right] ds \quad (11)$$

where $p_t = \nabla \tilde{\Phi}(\hat{G}_{t-1})$, $\hat{G}_{-i} = \max_{j \neq i} \hat{G}_{t-1,j} + Z_j$ and $\text{supp}(p_t) = \{i : p_{t,i} > 0\}$.

Note that each summand in the divergence penalty expression above involves an integral of the density function of the distribution \mathcal{D} over an interval. The main idea to control the divergence penalty for a bounded support distribution is to truncate the interval at the end of the support. For points that are close to the end of the support, we bound the integral by the product of the bound on the density and the interval length. For points that are far from the end of the support, we bound the integral through the hazard rate as was done by Abernethy et al. [2015].

For a general continuous random variable Z with bounded density, bounded support, we first shift it (which obviously does not change the action choice i_t and hence the expected regret) and scale it so that the support is a subset of $[0, 1]$ with $\inf\{z : F(z) = 0\} = 0$ and $\inf\{z : F(z) = 1\} = 1$ where F denotes the CDF of Z . A benefit of this normalization is that the expectation of the random variable becomes non-negative so the underestimation penalty is guaranteed to be non-positive. After scaling, we assume that the bound on the density is L . We consider the perturbation ηZ where $\eta > 0$ is a tuning parameter. Write $F_\eta(x)$ and $f_\eta(x)$ to denote the CDF and PDF of the scaled random variable ηZ respectively. If F is strictly increasing, we know that F^{-1} exists. If not, define $F^{-1}(y) = \inf\{z : F(z) = y\}$. Elementary calculation gives the following useful facts:

$$F_\eta(z) = F\left(\frac{z}{\eta}\right), f_\eta(z) = \frac{f\left(\frac{z}{\eta}\right)}{\eta}, F_\eta^{-1}(y) = \eta F^{-1}(y).$$

Theorem 3.2. (Divergence Penalty Control, Bounded Support) *The divergence penalty in the GBPA regret bound using the perturbation ηZ , where Z is drawn from a bounded support distribution satisfying the conditions above, can be upper bounded, for any $\epsilon > 0$, by*

$$NL \left(\frac{1}{2\eta\epsilon} + 1 - F^{-1}(1 - \epsilon) \right).$$

The regret bound for the uniform distribution is now an easy corollary.

Corollary 3.3. (Regret Bound for Uniform) *For GBPA run with a stochastic smoothing using an appropriately scaled $[0, 1]$ uniform perturbation, the expected regret can be upper bounded by $3(NT)^{2/3}$.*

For a general perturbation with bounded support and bounded density, the rate at which $1 - F^{-1}(1 - \epsilon)$ goes to 0 as $\epsilon \rightarrow 0$ can vary but we can always guarantee sublinear expected regret.

Corollary 3.4. (Asymptotic Regret Bound for Bounded Support) *For stochastically smoothed GBPA using general continuous random variable Z with bounded density and bounded support contained in $[0, 1]$, the expected regret grows sublinearly, i.e.,*

$$\lim_{T \rightarrow \infty} \frac{\mathbb{E} \text{Regret}_T}{T} = 0.$$

4 Perturbations with Unbounded Support

Unlike perturbations with bounded support, perturbations with unbounded support (on the right) do have non-zero right tail probabilities, ensuring that $p_{t,i} > 0$ always. However, the tail behavior may be such that the hazard rate is unbounded. Still, under mild assumptions, perturbations with unbounded support (on the right) can also be shown to have near optimal expected regret in T , using the notion of *generalized hazard rate* that we now introduce.

4.1 Generalized Hazard Rate

We already know how to control the underestimation and overestimation penalties via Lemma 2.2. So our main focus will be to control the divergence penalty. Towards this end, we define the generalized hazard rate for a continuous random variable Z with support unbounded on the right, parameterized by $\alpha \in [0, 1)$, as

$$h_\alpha(z) := \frac{f(z)|z|^\alpha}{(1 - F(z))^{1-\alpha}}, \quad (12)$$

where $f(z)$ and $F(z)$ denotes the PDF and CDF of Z respectively. Note that by setting $\alpha = 0$ we recover the standard hazard rate.

One of the main results of this paper is the following. Note that it includes the result (Lemma 4.3) of Abernethy et al. [2015] as a special case.

Theorem 4.1. (Divergence Penalty Control via Generalized Hazard Rate) *Let $\alpha \in [0, 1)$. Suppose we have $\forall z \in \mathbb{R}, h_\alpha(z) \leq C$. Then,*

$$\mathbb{E}[D_{\Phi}(\hat{G}_t, \hat{G}_{t-1}) | i_{1:t-1}] \leq \frac{2C}{1-\alpha} \times N.$$

A regret bound now easily follows.

Theorem 4.2. (Regret Bound via Generalized Hazard Rate) *Suppose we use a stochastic smoothing with a perturbation distribution whose generalized hazard rate is bounded: $h_\alpha(x) \leq C, \forall x \in \mathbb{R}$ for some $\alpha \in [0, 1)$, and*

$$\mathbb{E}_{Z_1, \dots, Z_N}[\max_i Z_i] - \mathbb{E}[Z_1] \leq Q(N),$$

where $Q(N)$ is some function of N . Then, the expected regret of GBPA is no greater than

$$2 \times \left(\frac{2C}{1-\alpha}\right)^{1/(2-\alpha)} \times (NT)^{1/(2-\alpha)} \times Q(N)^{(1-\alpha)/(2-\alpha)}.$$

In particular, this implies that the algorithm has sublinear expected regret.

4.2 Gaussian Perturbation

In this section we prove that GBPA with the standard Gaussian perturbation incurs a near optimal expected regret in both N and T . Let $F(z)$ and $f(z)$ denote the CDF and PDF of standard Gaussian distribution.

Lemma 4.3 (Baricz [2008]). *For standard Gaussian random variable, we have*

$$z < \frac{f(z)}{1 - F(z)} < \frac{z}{2} + \frac{\sqrt{z^2 + 4}}{2}.$$

This lemma together with example 2.6 in Thomas [1971] show that the hazard rate of a standard Gaussian random variable increases monotonically to infinity. However, we can still bound the generalized hazard rate for strictly positive α .

Lemma 4.4. (Generalized Hazard Bound for Gaussian) *For any $\alpha \in (0, 1)$, we have*

$$\frac{f(z)|z|^\alpha}{(1 - F(z))^{1-\alpha}} \leq C_1$$

where $C_1 = \frac{2}{\alpha}$.

The bounded generalized hazard rate shown in the above lemma can be used to control the divergence penalty. Combined with other knowledge of the standard Gaussian random variable we are able to give a bound on the expected regret.

Corollary 4.5. *The expected regret of GBPA with standard Gaussian random variable as perturbation has an expected regret at most*

$$2(C_1 C_2 N T)^{1/(2-\alpha)} (\sqrt{2 \log N})^{(1-\alpha)/(2-\alpha)}$$

where $C_1 = \frac{2}{\alpha}$, $C_2 = \frac{2}{1-\alpha}$, and $\alpha \in (0, 1)$.

It remains to optimally tune α in the above bound. Note the tuning parameter α appears only in the analysis, not in the algorithm.

Theorem 4.6. (Regret Bound for Gaussian) *The expected regret of GBPA with standard Gaussian random variable as perturbation has an expected regret at most*

$$96\sqrt{NT} \times N^{1/\log T} \sqrt{\log N} \log T$$

for $T > 4$. If we assume that $T > N$, the expected regret can be upper bounded by

$$278\sqrt{NT} \times \sqrt{\log N} \log T.$$

4.3 Sufficient Condition for Near Optimal Regret

In Section 4.1 we showed that if the generalized hazard rate of a distribution is bounded, the expected regret of the GBPA can be controlled. In this section, we are going to prove that under reasonable assumptions on the distribution of the perturbation, the FTPL enjoys near optimal expected regret.

Assumptions (a)-(c). Before we proceed, let us formally state our assumptions on the distributions we will consider. The distribution needs to (a) be continuous and has bounded density (b) has finite expectation (c) has support unbounded in the $+\infty$ direction.

Note that if the expectation of the random perturbation is negative, we shift it so that the expectation is zero. Hence the underestimation penalty is non-positive. In addition to the assumptions we have made above, we make another assumption on the eventual monotonicity of the hazard rate.

Assumption (d) $h_0(z) = \frac{f(z)}{1-F(z)}$ is eventually monotone.

“Eventually monotone” means that $\exists z_0 \geq 0$ such that if $z > z_0$, $\frac{f(z)}{1-F(z)}$ is non-decreasing or non-increasing. This assumption might appear hard to check, but numerous theorems are available to establish the monotonicity of hazard rate, which is much stronger than what we are assuming here. For example, see Theorem 2.4 in Thomas [1971], Theorem 2 and Theorem 4 in Chechile [2003], Chechile [2009]. In fact, most natural distributions do satisfy this assumption [Bagnoli and Bergstrom, 2005].

Before we proceed, we mention a standard classification of random variables into two classes based on their tail property.

Definition 4.7 (see, for example, Foss et al. [2009]). *A function $f(z) \geq 0$ is said to be heavy-tailed if and only if*

$$\lim_{z \rightarrow \infty} \sup f(z) e^{\lambda z} = \infty \quad \text{for all } \lambda > 0.$$

A distribution with CDF $F(z)$ and $\bar{F}(z) = 1 - F(z)$ is said to be heavy-tailed if and only if $\bar{F}(z)$ is heavy-tailed. If the distribution is not heavy-tailed, we say that it is light-tailed.

It turns out that under assumptions (a)-(d), if the distribution is also heavy-tailed, then the hazard rate itself is bounded. If the distribution is light-tailed, we need an additional assumption on the eventual monotonicity of a function similar to generalized hazard rate to ensure the boundedness of the generalized hazard rate. But before we state and prove the main results, we introduce some functions and prove an intermediate lemma that will be useful to prove the main results.

Define $R(z) = -\log \bar{F}(z)$ so that we have $\bar{F}(z) = e^{-R(z)}$ and $R'(z) = \frac{f(z)}{\bar{F}(z)} = h_0(z)$.

Lemma 4.8. *Under assumptions (a)-(d), we have*

$$\overline{F}(z)e^{\lambda z} \text{ is eventually monotone } \forall \lambda > 0.$$

We are finally ready to present the main results in this section.

Theorem 4.9. (Heavy Tail Implies Bounded Hazard) *Under assumptions (a) - (d), if the distribution is also heavy-tailed, then the hazard rate is bounded, i.e.,*

$$\sup_z \frac{f(z)}{\overline{F}(z)} < \infty.$$

Unlike heavy-tailed distributions, the hazard rate of light-tailed distributions might be unbounded. However, it turns out that if we make an additional assumption on the eventual monotonicity of a function similar to the generalized hazard rate, we can still guarantee the boundedness of the generalized hazard rate.

Assumption (e) $\exists \delta \in (0, 1]$ such that $\frac{f(z)}{(1 - F(z))^{1-\delta}}$ is eventually monotone.

Theorem 4.10. (Light Tail Implies Bounded Generalized Hazard) *Under assumptions (a) - (e), if the distribution is also light-tailed, then for any $\alpha \in (\delta, 1)$, the generalized hazard rate $h_\alpha(z)$ is bounded, i.e.,*

$$\sup_z \frac{f(z)|z|^\alpha}{(\overline{F}(z))^{1-\alpha}} < \infty.$$

Combining the above result with control of the divergence penalty gives us the following corollary.

Corollary 4.11. *Under assumptions (a)-(e), if the distribution is also light-tailed, the expected regret of GBPA with perturbations drawn from that distribution is, for any $\alpha \in (\delta, 1)$ and $\xi > 0$,*

$$O\left((TN)^{1/(2-\alpha)}N^\xi\right).$$

In particular, if assumption (e) holds for any $\delta \in (0, 1)$, then the expected regret of GBPA is $O\left((TN)^{1/2+\epsilon}\right)$ for any $\epsilon > 0$, i.e., it is near optimal in both N and T .

Next we consider a family of light-tailed distributions that do not have a bounded hazard rate.

Definition 4.12. *The exponential power (or generalized normal) family of distributions, denoted as \mathcal{D}_β where $\beta > 1$, is defined via the cdf*

$$f_\beta(z) = C_\beta e^{-z^\beta}, \quad z \geq 0.$$

The next theorem shows that GBPA with perturbations from this family of distributions enjoys near optimal expected regret in both N and T .

Theorem 4.13. (Regret Bound for Power Exponential Family) $\forall \beta > 1$, *the expected regret of GBPA with perturbations drawn from \mathcal{D}_β is, for any $\epsilon > 0$, $O\left((TN)^{1/2+\epsilon}\right)$.*

5 Conclusion and Future Work

Previous work on providing regret guarantees for FTPL algorithms in the adversarial multi-armed bandit setting required a bounded hazard rate condition. We have shown how to go beyond the hazard rate condition but a number of questions remain open. For example, what if we use FTPL with perturbations from discrete distributions such as Bernoulli distribution? In the full information setting [Devroye et al. \[2013\]](#) and [Van Erven et al. \[2014\]](#) have considered random walk perturbation and dropout perturbation, both leading to minimax optimal regret. But to the best of our knowledge those distributions have not been analyzed in the adversarial multi-armed bandit problem.

An unsatisfactory aspect of even the tightest bounds for FTPL algorithms from existing work, including ours, is that they never reach the minimax optimal $O(\sqrt{NT})$ bound. They come very close to

it: up to logarithmic factors. It is known that FTRL algorithms, using the negative Tsallis entropy as the regularizer, can achieve the optimal bound [Audibert and Bubeck, 2009, Audibert et al., 2011, Abernethy et al., 2015]. Is there a perturbation that can achieve the optimal bound?

We only considered multi-armed bandits in this work. There has been some interest in using FTPL algorithms for combinatorial bandit problems (see, for example, Neu and Bartók [2013]). In future work, it will be interesting to extend our analysis to combinatorial bandit problems.

Acknowledgments. We thank Jacob Abernethy and Chansoo Lee for helpful discussions. We acknowledge the support of NSF under CAREER grant IIS-1452099.

References

- Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *COLT*, pages 807–823, 2014.
- Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems 28*, pages 2188–2196, 2015.
- Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.
- Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. In *COLT*, 2011.
- Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM J. Comput.*, 32:48–77, 2002.
- Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. *Economic Theory*, 26(2):445–469, 2005.
- Árpád Baricz. Mills’ ratio: Monotonicity patterns and functional inequalities. *J. Math. Anal. Appl.*, 340(2):1362–1370, 2008.
- Dimitri P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973. ISSN 0022-3239.
- Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.
- Richard A. Chechile. Mathematical tools for hazard function analysis. *J. Math. Psychol.*, 47:478–494, 2003.
- Richard A. Chechile. Corrigendum to: mathematical tools for hazard function analysis [j. math. psychol. 47 (2003) 478494]. *J. Math. Psychol.*, 53:298–299, 2009.
- Luc Devroye, Gábor Lugosi, and Gergely Neu. Prediction by random-walk perturbation. In *Conference on Learning Theory*, pages 460–473, 2013.
- Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An Introduction to Heavy-tailed and Subexponential Distributions*. Springer, 2009.
- J. Hannan. Approximation to bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games, volume III*, pages 97–139, 1957.
- Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.
- Jussi Kujala and Tapio Elomaa. On following the perturbed leader in the bandit setting. In *Algorithmic Learning Theory*, pages 371–385. Springer, 2005.

Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 234–248. Springer, 2013.

Jan Poland. FPL analysis for adaptive bandits. In Oleg B. Lupanov, Oktay M. Kasim-Zade, Alexander V. Chaskin, and Kathleen Steinhöfel, editors, *Stochastic Algorithms: Foundations and Applications: Third International Symposium, SAGA 2005, Moscow, Russia, October 20-22, 2005. Proceedings*, pages 58–69. Springer Berlin Heidelberg, 2005.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5): 527–535, 1952.

Ewart A. C. Thomas. Sufficient conditions for monotone hazard rate an application to latency-probability curves. *J. Math. Psychol.*, 8:303–332, 1971.

Tim Van Erven, Wojciech Kotłowski, and Manfred K Warmuth. Follow the leader with dropout perturbations. In *COLT*, 2014.

A Proofs

A.1 Proof of Lemma 2.1

Proof. First, note that the regret, by definition, is

$$\text{Regret}_T = \Phi(G_T) - \sum_{t=1}^T \langle \mathbf{e}_{i_t}, g_t \rangle.$$

Under an oblivious adversary, only the summation on the right hand side is random. Moreover $\mathbb{E}[\langle \mathbf{e}_{i_t}, g_t \rangle | i_{1:t-1}] = \langle p_t, g_t \rangle$. This proves the claim in (4).

From (2), we know that $\mathbb{E}[\langle p_t, \hat{g}_t \rangle | i_{1:t-1}] = \langle p_t, g_t \rangle$ even if some entries in p_t might be zero. Therefore, we have

$$\mathbb{E}\text{Regret}_T = \Phi(G_T) - \mathbb{E} \left[\sum_{t=1}^T \langle p_t, \hat{g}_t \rangle \right]. \quad (13)$$

From (3), we know that $G_T \leq \mathbb{E}[\hat{G}_T]$. This implies

$$\Phi(G_T) \leq \Phi(\mathbb{E}[\hat{G}_T]) \leq \mathbb{E}[\Phi(\hat{G}_T)], \quad (14)$$

where the first inequality is because $G \succeq G' \Rightarrow \Phi(G) \geq \Phi(G')$, and the second inequality is due to the convexity of Φ . Plugging (14) into (13) yields

$$\mathbb{E}\text{Regret}_T \leq \mathbb{E} \left[\Phi(\hat{G}_T) - \sum_{t=1}^T \langle p_t, \hat{g}_t \rangle \right]. \quad (15)$$

Now considering the quantity inside the expectation above and recalling the definition of Bregman divergence

$$D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) = \tilde{\Phi}(\hat{G}_t) - \tilde{\Phi}(\hat{G}_{t-1}) - \left\langle \nabla \tilde{\Phi}(\hat{G}_{t-1}), \hat{G}_t - \hat{G}_{t-1} \right\rangle$$

we get,

$$\begin{aligned} \Phi(\hat{G}_T) - \sum_{t=1}^T \langle p_t, \hat{g}_t \rangle &= \Phi(\hat{G}_T) - \sum_{t=1}^T \left\langle \nabla \tilde{\Phi}(\hat{G}_{t-1}), \hat{g}_t \right\rangle \\ &= \Phi(\hat{G}_T) - \sum_{t=1}^T \left\langle \nabla \tilde{\Phi}(\hat{G}_{t-1}), \hat{G}_t - \hat{G}_{t-1} \right\rangle \\ &= \Phi(\hat{G}_T) + \sum_{t=1}^T \left(D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) + \tilde{\Phi}(\hat{G}_{t-1}) - \tilde{\Phi}(\hat{G}_t) \right) \\ &= \Phi(\hat{G}_T) + \tilde{\Phi}(\hat{G}_0) - \tilde{\Phi}(\hat{G}_T) + \sum_{t=1}^T D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}). \end{aligned} \quad (16)$$

The proof ends by plugging in (16) into (15) and noting that $\tilde{\Phi}(\hat{G}_0) = \tilde{\Phi}(0)$ is not random. \square

A.2 Proof of Lemma 2.2

Proof. We have,

$$\begin{aligned}\Phi(G) + \mathbb{E}[Z_1] &= \max_i G_i + \mathbb{E}[Z_i] = \max_i (G_i + \mathbb{E}[Z_i]) \\ &\leq \mathbb{E}[\max_i (G_i + Z_i)] = \tilde{\Phi}(G) \\ &\leq \mathbb{E}[\max_i G_i + \max_i Z_i] = \max_i G_i + \mathbb{E}[\max_i Z_i] = \Phi(G) + \mathbb{E}[\max_i Z_i].\end{aligned}$$

Noting that $\mathbb{E}[\max_i Z_i] \leq EMAX(N)$ finishes the proof. \square

A.3 Proof of Lemma 3.1

Proof. To reduce clutter, we drop the time subscripts: we use \hat{G} to denote the cumulative estimate \hat{G}_{t-1} , \hat{g} to denote the marginal estimate $\hat{g}_t = \hat{G}_t - \hat{G}_{t-1}$, p to denote p_t , and g to denote the true loss g_t . Note that by definition of Framework 1, \hat{g} is a sparse vector with one non-zero and non-positive coordinate $\hat{g}_{i_t} = g_{i_t}/p_{i_t} = -|g_{i_t}/p_{i_t}|$. Moreover, conditioned on $i_{1:t-1}$, i_t takes value i with probability p_i . For any $i \in \text{supp}(p)$, let

$$h_i(r) = D_{\tilde{\Phi}}(\hat{G} - r\mathbf{e}_i, \hat{G}),$$

so that $h'_i(r) = -\nabla_i \tilde{\Phi}(\hat{G} - r\mathbf{e}_i) + \nabla_i \tilde{\Phi}(\hat{G})$ and $h''_i(r) = \nabla_{ii}^2 \tilde{\Phi}(\hat{G} - r\mathbf{e}_i)$. Now we write:

$$\begin{aligned}\mathbb{E}[D_{\tilde{\Phi}}(\hat{G} + \hat{g}, \hat{G}) | i_{1:t-1}] &= \sum_{i \in \text{supp}(p)} p_i D_{\tilde{\Phi}}(\hat{G} + g_i/p_i \mathbf{e}_i, \hat{G}) = \sum_{i \in \text{supp}(p)} p_i D_{\tilde{\Phi}}(\hat{G} - |g_i/p_i| \mathbf{e}_i, \hat{G}) \\ &= \sum_{i \in \text{supp}(p)} p_i h_i(|g_i/p_i|) = \sum_{i \in \text{supp}(p)} p_i \int_0^{|g_i/p_i|} \int_0^s h''_i(r) dr ds \\ &= \sum_{i \in \text{supp}(p)} p_i \int_0^{|g_i/p_i|} \int_0^s \nabla_{ii}^2 \tilde{\Phi}(\hat{G} - r\mathbf{e}_i) dr ds \\ &= \sum_{i \in \text{supp}(p)} p_i \int_0^{|g_i/p_i|} \int_0^s \mathbb{E}_{\hat{G}-i} f(\hat{G}_{-i} - \hat{G}_i + r) dr ds \\ &= \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{|g_i/p_i|} \mathbb{E}_{\hat{G}-i} \left[\int_0^s f(\hat{G}_{-i} - \hat{G}_i + r) dr \right] ds.\end{aligned}$$

\square

A.4 Proof of Theorem 3.2

Proof. From Lemma 3.1, we have, with $\hat{G}_{-i} = \max_{j \neq i} \hat{G}_{t-1,j} + \eta Z_j$,

$$\begin{aligned}
& \mathbb{E} \left[D_{\hat{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) | i_{1:t-1} \right] \\
&= \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{\left| \frac{g_{t,i}}{p_{t,i}} \right|} \mathbb{E}_{\hat{G}_{-i}} \left[\int_0^s f_{\eta}(\hat{G}_{-i} - \hat{G}_{t-1,i} + r) dr \right] ds \\
&= \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{\left| \frac{g_{t,i}}{p_{t,i}} \right|} \mathbb{E}_{\hat{G}_{-i}} \left[\int_{\hat{G}_{-i} - \hat{G}_{t-1,i}}^{\hat{G}_{-i} - \hat{G}_{t-1,i} + s} f_{\eta}(z) dz \right] ds \\
&= \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{\left| \frac{g_{t,i}}{p_{t,i}} \right|} \left(\mathbb{E}_{\hat{G}_{-i}} \left[\underbrace{\int_{[\hat{G}_{-i} - \hat{G}_{t-1,i}, \hat{G}_{-i} - \hat{G}_{t-1,i} + s] \setminus [F_{\eta}^{-1}(1-\epsilon), \eta]} f_{\eta}(z) dz}_{(I)} \right] \right. \\
&\quad \left. + \underbrace{\int_{[F_{\eta}^{-1}(1-\epsilon), \eta]} f_{\eta}(z) dz}_{(II)} \right) ds. \tag{17}
\end{aligned}$$

We bound the two integrals above differently. For the first integral, we add the restriction $f_{\eta}(z) > 0$ by intersecting the integral interval with the support of the function $f_{\eta}(z)$, denoted as $I_{f_{\eta}(z)}$ so that $1 - F_{\eta}(z)$ is not 0 on the interval to be integrated. Thus, we get,

$$\begin{aligned}
(I) &= \int_{([\hat{G}_{-i} - \hat{G}_{t-1,i}, \hat{G}_{-i} - \hat{G}_{t-1,i} + s] \setminus [F_{\eta}^{-1}(1-\epsilon), \eta]) \cap I_{f_{\eta}(z)}} f_{\eta}(z) dz \\
&= \int_{([\hat{G}_{-i} - \hat{G}_{t-1,i}, \hat{G}_{-i} - \hat{G}_{t-1,i} + s] \setminus [F_{\eta}^{-1}(1-\epsilon), \eta]) \cap I_{f_{\eta}(z)}} (1 - F_{\eta}(z)) \cdot \frac{f_{\eta}(z)}{1 - F_{\eta}(z)} dz \\
&\leq \int_{([\hat{G}_{-i} - \hat{G}_{t-1,i}, \hat{G}_{-i} - \hat{G}_{t-1,i} + s] \setminus [F_{\eta}^{-1}(1-\epsilon), \eta]) \cap I_{f_{\eta}(z)}} (1 - F_{\eta}(z)) \cdot \frac{L}{\eta \epsilon} \\
&\leq (1 - F_{\eta}(\hat{G}_{-i} - \hat{G}_{t-1,i})) \frac{sL}{\eta \epsilon}. \tag{18}
\end{aligned}$$

The first inequality holds because $f_{\eta}(z) \leq L/\eta$ and $(1 - F_{\eta}(z)) \geq \epsilon$ on the set of z 's over which we are integrating. The second inequality holds because on the set under consideration $1 - F_{\eta}(z) \leq 1 - F_{\eta}(\hat{G}_{-i} - \hat{G}_{t-1,i})$ and the measure of the set is at most s .

For the second integral, we use the bound $f_{\eta}(z) \leq L/\eta$ again to get,

$$(II) = \int_{[F_{\eta}^{-1}(1-\epsilon), \eta]} f_{\eta}(z) dz \leq \frac{L}{\eta} \cdot (\eta - F_{\eta}^{-1}(1-\epsilon)). \tag{19}$$

Plugging in (18) and (19) into (17), we can bound the divergence penalty by,

$$\begin{aligned}
&\leq \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{\left| \frac{g_{t,i}}{p_{t,i}} \right|} \left(\mathbb{E}_{\hat{G}_{-i}} [1 - F_{\eta}(\hat{G}_{-i} - \hat{G}_{t-1,i})] \frac{sL}{\eta \epsilon} + \frac{L(\eta - F_{\eta}^{-1}(1-\epsilon))}{\eta} \right) ds \\
&= \sum_{i \in \text{supp}(p_t)} p_{t,i} \int_0^{\left| \frac{g_{t,i}}{p_{t,i}} \right|} \left(p_{t,i} \frac{sL}{\eta \epsilon} + L(1 - F^{-1}(1-\epsilon)) \right) ds \\
&= \sum_{i \in \text{supp}(p_t)} p_{t,i} \left(p_{t,i} \frac{L}{\eta \epsilon} \frac{g_{t,i}^2}{2p_{t,i}^2} + L(1 - F^{-1}(1-\epsilon)) \frac{|g_{t,i}|}{p_{t,i}} \right) \\
&\leq \sum_{i \in \text{supp}(p_t)} \left(\frac{L}{2\eta \epsilon} + L(1 - F^{-1}(1-\epsilon)) \right) \\
&\leq NL \left(\frac{1}{2\eta \epsilon} + 1 - F^{-1}(1-\epsilon) \right).
\end{aligned}$$

The second to last inequality holds because $|g_{t,i}| \leq 1$ and the last inequality holds because the sum over i is at most over all N arms. \square

A.5 Proof of Corollary 3.3

Proof. For $[0, 1]$ uniform distribution, we have $L = 1$, $F^{-1}(1 - \epsilon) = 1 - \epsilon$ so the divergence penalty is upper bounded by

$$NT\left(\frac{1}{2\eta\epsilon} + \epsilon\right).$$

If we let $\epsilon = \frac{1}{\sqrt{2\eta}}$, we can see that the divergence penalty is upper bounded by $NT\sqrt{\frac{2}{\eta}}$. Together with the overestimation penalty which is trivially bounded by η and a non-positive underestimation penalty, we see that the final regret bound is

$$NT\sqrt{\frac{2}{\eta}} + \eta.$$

Setting $\eta = (NT)^{2/3}$ gives the desired result. \square

A.6 Proof of Corollary 3.4

Proof. For a general distribution, let $\epsilon = \frac{1}{\sqrt{\eta}}$. Since the overestimation penalty is trivially bounded by η and the underestimation penalty is non-positive, the expected regret can be upper bounded by

$$LNT\left(\frac{1}{2\sqrt{\eta}} + 1 - F^{-1}\left(1 - \frac{1}{\sqrt{\eta}}\right)\right) + \eta.$$

Setting $\eta = (NT)^{2/3}$ we see that the expected regret can be upper bounded by

$$\left(\frac{L}{2} + 1\right)(NT)^{2/3} + LNT\left(1 - F^{-1}\left(1 - \frac{1}{\sqrt{\eta}}\right)\right).$$

Since

$$\lim_{T \rightarrow \infty} 1 - F^{-1}\left(1 - \frac{1}{\sqrt{\eta}}\right) = \lim_{\eta \rightarrow \infty} 1 - F^{-1}\left(1 - \frac{1}{\sqrt{\eta}}\right) = 1 - F^{-1}(1) = 0,$$

we conclude that

$$\lim_{T \rightarrow \infty} \frac{\mathbb{E}\text{Regret}_T}{T} = 0.$$

\square

A.7 Proof of Theorem 4.1

Proof. Because of the unbounded support of Z , $\text{supp}(p_t) = \{1, \dots, N\}$. Lemma 3.1 gives us:

$$\begin{aligned} \mathbb{E}[D_{\tilde{\Phi}}(\hat{G}_t, \hat{G}_{t-1}) | i_{1:t-1}] &= \sum_{i=1}^N p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} \mathbb{E}_{\tilde{G}_{-i}} \int_0^s f(\tilde{G}_{-i} - \hat{G}_{t-1,i} + r) dr ds \\ &= \sum_{i=1}^N p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} \mathbb{E}_{\tilde{G}_{-i}} \int_{\tilde{G}_{-i} - \hat{G}_{t-1,i}}^{\tilde{G}_{-i} - \hat{G}_{t-1,i} + s} f(z) dz ds \\ &\leq C \sum_{i=1}^N p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} \mathbb{E}_{\tilde{G}_{-i}} \int_{\tilde{G}_{-i} - \hat{G}_{t-1,i}}^{\tilde{G}_{-i} - \hat{G}_{t-1,i} + s} (1 - F(z))^{1-\alpha} |z|^{-\alpha} dz ds \\ &\leq C \sum_{i=1}^N p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} \mathbb{E}_{\tilde{G}_{-i}} (1 - F(\tilde{G}_{-i} - \hat{G}_{t-1,i}))^{1-\alpha} \int_{\tilde{G}_{-i} - \hat{G}_{t-1,i}}^{\tilde{G}_{-i} - \hat{G}_{t-1,i} + s} |z|^{-\alpha} dz ds. \end{aligned}$$

Since the function $f(z) = |z|^{-\alpha}$ is symmetric, monotonically decreasing as $|z| \rightarrow \infty$, we have

$$\int_{\tilde{G}_{-i} - \hat{G}_{t-1,i}}^{\tilde{G}_{-i} - \hat{G}_{t-1,i} + s} |z|^{-\alpha} dz \leq \int_{-s/2}^{s/2} |z|^{-\alpha} dz = \frac{2^\alpha}{1-\alpha} s^{1-\alpha}.$$

Also, note that $z^{1-\alpha}$ is concave. Hence, by Jensen's inequality,

$$\mathbb{E}_{\tilde{G}_{-i}}[(1 - F(\tilde{G}_{-i} - \hat{G}_{t-1,i}))^{1-\alpha}] \leq (\mathbb{E}_{\tilde{G}_{-i}}[1 - F(\tilde{G}_{-i} - \hat{G}_{t-1,i})])^{1-\alpha} = p_{t,i}^{1-\alpha}.$$

Hence,

$$\begin{aligned} \mathbb{E}[D_{\Phi}(\hat{G}_t, \hat{G}_{t-1}) | i_{1:t-1}] &\leq \frac{2^\alpha C}{1-\alpha} \sum_{i=1}^N p_{t,i} \int_0^{|g_{t,i}/p_{t,i}|} p_{t,i}^{1-\alpha} s^{1-\alpha} ds \\ &= \frac{2^\alpha C}{1-\alpha} \sum_{i=1}^N p_{t,i}^{2-\alpha} \int_0^{|g_{t,i}/p_{t,i}|} s^{1-\alpha} ds \\ &= \frac{2^\alpha C}{(1-\alpha)(2-\alpha)} \sum_{i=1}^N p_{t,i}^{2-\alpha} |g_{t,i}/p_{t,i}|^{2-\alpha} \\ &= \frac{2^\alpha C}{(1-\alpha)(2-\alpha)} \sum_{i=1}^N |g_{t,i}|^{2-\alpha} \\ &\leq \frac{2^\alpha C}{(1-\alpha)(2-\alpha)} N \leq \frac{2C}{1-\alpha} N. \end{aligned}$$

□

A.8 Proof of Theorem 4.2

Proof. The divergence penalty can be controlled through Theorem 4.1 once we have bounded generalized hazard rate. It remains to control the overestimation and underestimation penalty. By Lemma 2.2, they are at most $\mathbb{E}_{Z_1, \dots, Z_n}[\max_i Z_i]$ and $-\mathbb{E}[Z_1]$ respectively. Suppose we scale the perturbation Z by $\eta > 0$, i.e., we add ηZ_i to each coordinate. It is easy to see that $\mathbb{E}[\max_{i=1, \dots, n} \eta Z_i] = \eta \mathbb{E}[\max_{i=1, \dots, n} Z_i]$ and $\mathbb{E}[\eta Z_1] = \eta \mathbb{E}[Z_1]$. For the divergence penalty, observe that $F_\eta(t) = F(t/\eta)$ and thus $f_\eta(t) = \frac{1}{\eta} f(t/\eta)$. Hence, the constant in the assumption needs to scale by $\eta^{\alpha-1}$. Plugging new bounds for the scaled perturbations into Lemma 2.1 gives us

$$\mathbb{E}\text{Regret}_T \leq \eta^{\alpha-1} \frac{2C}{1-\alpha} \times NT + \eta Q(N).$$

Setting $\eta = (\frac{2CNT}{(1-\alpha)Q(N)})^{1/(2-\alpha)}$ finishes the proof. □

A.9 Proof of Lemma 4.4

Proof. Since the numerator of the left hand side is an even function of z , and the denominator is a decreasing function, and the inequality is trivially true when $z = 0$, it suffices to prove for $z > 0$, which we assume for the rest of the proof. From Lemma 4.3 we can derive that

$$\frac{f(z)}{1 - F(z)} < z + 1.$$

Therefore,

$$\begin{aligned} \frac{f(z)|z|^\alpha}{(1 - F(z))^{1-\alpha}} &\leq \frac{f(z)z^\alpha}{(\frac{f(z)}{z+1})^{1-\alpha}} = (f(z)z)^\alpha (z+1)^{1-\alpha} \\ &\leq f(z)^\alpha (z+1) \leq z f(z)^\alpha + 1 = \sqrt{\frac{1}{2\pi}} z e^{-\alpha z^2/2} + 1. \end{aligned}$$

Let $g(z) = z e^{-\alpha z^2/2}$, $g'(z) = (1 - \alpha z^2) e^{-\alpha z^2/2}$. Therefore $g(z)$ is maximized at $z^* = \sqrt{\frac{1}{\alpha}}$. Therefore,

$$\frac{f(z)|z|^\alpha}{(1 - F(z))^{1-\alpha}} \leq \sqrt{\frac{1}{2\pi}} z e^{-\alpha z^2/2} + 1 \leq \sqrt{\frac{1}{2\pi}} z^* + 1 \leq z^* + 1 = \sqrt{\frac{1}{\alpha}} + 1 \leq \frac{2}{\alpha}.$$

□

A.10 Proof of Corollary 4.5

Proof. It is known that for standard Gaussian random variable, we have $\mathbb{E}[Z_1] = 0$ and

$$\mathbb{E}_{Z_1, \dots, Z_n} [\max_i Z_i] \leq \sqrt{2 \log N}.$$

Plug in to Theorem 4.2 gives the result. \square

A.11 Proof of Theorem 4.6

Proof. From Corollary 4.5 we see that the expected regret can be upper bounded by

$$2(C_1 C_2 N T)^{1/(2-\alpha)} (\sqrt{2 \log N})^{(1-\alpha)/(2-\alpha)}$$

where $C_1 = \frac{2}{\alpha}$ and $C_2 = \frac{2}{1-\alpha}$. Note that

$$\begin{aligned} & 2(C_1 C_2 N T)^{1/(2-\alpha)} (\sqrt{2 \log N})^{(1-\alpha)/(2-\alpha)} \\ & \leq 4(C_1 C_2)^{1/(2-\alpha)} N^{1/(2-\alpha)} \sqrt{\log N}^{(1-\alpha)/(2-\alpha)} T^{1/(2-\alpha)} \\ & = 4N^{1/(2-\alpha)} \sqrt{\log N}^{(1-\alpha)/(2-\alpha)} T^{1/2} \times (C_1 C_2)^{1/(2-\alpha)} T^{\alpha/(4-2\alpha)} \\ & \leq 4N^{1/2} N^{\alpha/(4-2\alpha)} \sqrt{\log N} T^{1/2} \times \left(\frac{4}{\alpha(1-\alpha)}\right)^{1/(2-\alpha)} T^{\alpha/(4-2\alpha)} \\ & \leq 4N^{1/2} N^{\alpha} \sqrt{\log N} T^{1/2} \times \frac{4T^{\alpha}}{\alpha(1-\alpha)} \\ & \leq 16\sqrt{NT} N^{\alpha} \sqrt{\log N} \times \frac{T^{\alpha}}{\alpha(1-\alpha)}. \end{aligned}$$

\square

If we let $\alpha = \frac{1}{\log T}$, then $T^{\alpha} = T^{1/\log T} = e < 3$. Then, we have, for $T > 4$,

$$\frac{T^{\alpha}}{\alpha(1-\alpha)} \leq \frac{3 \log T}{1 - \frac{1}{\log T}} = \frac{3 \log^2 T}{\log T - 1} \leq 6 \log T.$$

Putting things together finishes the proof.

A.12 Proof of Lemma 4.8

Proof. Let $g(z) = \overline{F}(z)e^{\lambda z}$, then $g'(z) = e^{\lambda z} \overline{F}(z)(\lambda - \frac{f(z)}{F(z)})$. Since $\frac{f(z)}{F(z)}$ is eventually monotone by assumption (d), $g'(z)$ is eventually positive, negative or zero. The lemma immediately follows. \square

A.13 Proof of Theorem 4.9

Proof. If the distribution is heavy-tailed, we have

$$\lim_{z \rightarrow \infty} \sup \overline{F}(z)e^{\lambda z} = \infty \quad \text{for all } \lambda > 0.$$

By Lemma 4.8, we can erase the supremum operator and just write

$$\lim_{z \rightarrow \infty} \overline{F}(z)e^{\lambda z} = \infty \quad \text{for all } \lambda > 0.$$

Hence,

$$\lim_{z \rightarrow \infty} \overline{F}(z)e^{\lambda z} = \lim_{x \rightarrow \infty} e^{-R(z)+\lambda z} = \infty \text{ for all } \lambda > 0 \Rightarrow \lim_{z \rightarrow \infty} \sup \frac{R(z)}{z} = 0.$$

Note that $R'(z) = \frac{f(z)}{F(z)}$, which is eventually monotone by assumption. Therefore, we can conclude that

$$\lim_{z \rightarrow \infty} \sup R'(z) < \infty \Rightarrow \sup_z \frac{f(z)}{F(z)} < \infty.$$

\square

A.14 Proof of Theorem 4.10

Proof. If the distribution is light-tailed, we have

$$\lim_{z \rightarrow \infty} \overline{F}(z)e^{\lambda^* z} < \infty \quad \text{for some } \lambda^* > 0. \quad (20)$$

This immediately implies that

$$\lim_{z \rightarrow +\infty} \overline{F}(z)^a z^b = 0 \quad \forall a, b > 0. \quad (21)$$

Consider $\lim_{z \rightarrow \infty} \frac{f(z)}{\overline{F}(z)} = \lim_{z \rightarrow \infty} R'(z)$. If $\lim_{z \rightarrow \infty} R'(z) < \infty$ we can immediately conclude that $\sup_z \frac{f(z)}{1-F(z)} < \infty$. If $\lim_{z \rightarrow \infty} R'(z) = \infty$ instead, note that

$$\lim_{z \rightarrow \infty} \int_{-z}^z R'(t)e^{-\delta R(t)} dt = -\frac{1}{\delta} e^{-\delta R(z)} \Big|_{z=-\infty}^{z=+\infty} = \frac{1}{\delta} < \infty.$$

Moreover, since $\lim_{z \rightarrow \infty} R'(z) = \infty$, $R'(z)e^{-\delta R(z)}$ is strictly positive for all $z > z_0$ for some z_0 . Furthermore, $R'(z)e^{-\delta R(z)} = \frac{f(z)}{(\overline{F}(z))^{1-\delta}}$ is eventually monotone by assumption (e),

Therefore, we can conclude that

$$\lim_{z \rightarrow \infty} R'(z)e^{-\delta R(z)} = \frac{f(z)}{(\overline{F}(z))^{1-\delta}} = 0.$$

$\forall \alpha \in (\delta, 1)$, from Equation (21) we have $\lim_{z \rightarrow +\infty} z^\alpha \overline{F}(z)^{\alpha-\delta} = 0$, so

$$\lim_{z \rightarrow +\infty} \frac{f(z)z^\alpha}{(\overline{F}(z))^{1-\alpha}} = \lim_{z \rightarrow +\infty} \frac{f(z)}{\overline{F}(z)^{1-\delta}} \times z^\alpha \overline{F}(z)^{\alpha-\delta} = 0.$$

and hence

$$\sup_z \frac{f(z)z^\alpha}{(1-F(z))^{1-\alpha}} < \infty \quad \forall \alpha \in (\delta, 1).$$

□

A.15 Proof of Corollary 4.11

Proof. For a light-tailed distribution \mathcal{D} , we have

$$\lim_{z \rightarrow \infty} \overline{F}_{\mathcal{D}}(z)e^{\lambda^* z} < \infty \quad \text{for some } \lambda^* > 0.$$

This implies that

$$\overline{F}_{\mathcal{D}}(z) \leq Ce^{-\lambda^* z} \text{ for some } C > 0, z > z_0.$$

Let random variable Z follows distribution \mathcal{D} . Since Z might take negative values, we define a new distribution \mathcal{D}' that only takes non-negative value by

$$f_{\mathcal{D}'}(z) = \begin{cases} \frac{1}{p_{\mathcal{D}+}} f_{\mathcal{D}}(z) & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

where $p_{\mathcal{D}+} = \mathbb{P}(Z \geq 0) > 0$ by right unbounded support assumption. Clearly, with this definition of \mathcal{D}' we see that $\mathbb{E}_{Z_1, \dots, Z_N \sim \mathcal{D}}[\max_i Z_i] \leq \mathbb{E}_{Z_1, \dots, Z_N \sim \mathcal{D}'}[\max_i Z_i]$ and for $z > z_0$, we have $\overline{F}_{\mathcal{D}'}(z) = \frac{\overline{F}_{\mathcal{D}}(z)}{p_{\mathcal{D}+}} \leq C'e^{-\lambda^* z}$

where $C' = \frac{C}{p_{D+}}$. Note that

$$\begin{aligned}
\mathbb{E}_{Z_1, \dots, Z_N \sim \mathcal{D}}[\max_i Z_i] &\leq \mathbb{E}_{Z_1, \dots, Z_N \sim \mathcal{D}'}[\max_i Z_i] \\
&= \int_0^\infty \mathbb{P}(\max_i Z_i > x) dx \\
&\leq u + \int_u^\infty \mathbb{P}(\max_i Z_i > z) dz \\
&\leq u + N \int_u^\infty \mathbb{P}(Z_i > z) dz \\
&\leq u + N \int_u^\infty C' e^{-\lambda^* z} dz \quad \text{assuming } u > z_0 \\
&= u + \frac{C' N}{\lambda^*} e^{-\lambda^* u}.
\end{aligned}$$

If we let $u = \frac{\log(N)}{\lambda^*}$, obviously $u > z_0$ if N is sufficiently large. Thus, we see that

$$\mathbb{E}_{Z_1, \dots, Z_N \sim \mathcal{D}}[\max_i Z_i] \leq \frac{\log(N)}{\lambda^*} + C' = O(N^\xi) \quad \forall \xi > 0. \quad (22)$$

From Theorem 4.10 we see that $\forall \alpha \in (\delta, 1)$,

$$\frac{f(z)z^\alpha}{(1 - F(z))^{1-\alpha}} \leq C_\alpha \quad \forall z \in \mathbb{R}. \quad (23)$$

Plug 22 and 23 into Theorem 4.2 gives the desired result. \square

A.16 Proof of Corollary 4.13

Proof. By Corollary 4.11 we only need to check that assumptions (a)-(d) hold for distribution \mathcal{D}_β , exponential power family is light-tailed, and assumption (e) also holds for any $\delta \in (0, 1)$. By observing the density function f_β we can trivially see that assumptions (a)-(c) hold and that the subbotin family is light-tailed. Therefore, define

$$g_{\delta, \beta}(z) = \frac{f_\beta(z)}{(\overline{F}_\beta(z))^{1-\delta}} = \frac{f_\beta(z)}{(1 - F_\beta(z))^{1-\delta}},$$

it suffices to show that $\forall \delta \in [0, 1)$, $g_{\delta, \beta}(z)$ is eventually monotone. Note that

$$\begin{aligned}
g'_{\delta, \beta}(z) &= \frac{f'_\beta(z)(1 - F_\beta(z))^{1-\delta} + (1 - \delta)(1 - F_\beta(z))^{-\delta} f_\beta^2(z)}{(1 - F_\beta(z))^{2-2\delta}} \\
&= \frac{C_\beta^2 e^{-z^\beta}}{(1 - F_\beta(z))^{2-\delta}} \times \left((1 - \delta)e^{-z^\beta} - \beta z^{\beta-1} \int_z^\infty e^{-t^\beta} dt \right).
\end{aligned}$$

It further suffices to show that

$$m_{\delta, \beta}(z) = (1 - \delta)e^{-z^\beta} - \beta z^{\beta-1} \int_z^\infty e^{-t^\beta} dt$$

is eventually non-negative or non-positive $\forall \beta > 1, \delta \in [0, 1)$. Note that since $\beta > 1$,

$$\beta z^{\beta-1} \int_z^\infty e^{-t^\beta} dt = \int_z^\infty \beta t^{\beta-1} e^{-t^\beta} dt < \int_z^\infty \beta t^{\beta-1} e^{-t^\beta} dt = e^{-z^\beta}. \quad (24)$$

Therefore, $m_{0, \beta}(z) > 0$ for all $z \geq 0$, i.e., the hazard rate is always increasing and assumption (d) is satisfied. Now, we are left to show that $m_{\delta, \beta}(z)$ is eventually non-negative or non-positive for any

$\delta \in (0, 1)$. Note that

$$\begin{aligned}
\beta z^{\beta-1} \int_z^\infty e^{-t^\beta} dt &= \beta \left(\frac{z}{z+1} \right)^{\beta-1} (z+1)^{\beta-1} \int_z^\infty e^{-t^\beta} dt \\
&\geq \beta \left(\frac{z}{z+1} \right)^{\beta-1} (z+1)^{\beta-1} \int_z^{z+1} e^{-t^\beta} dt \\
&\geq \left(\frac{z}{z+1} \right)^{\beta-1} \int_z^{z+1} \beta t^{\beta-1} e^{-t^\beta} dt \\
&= \left(\frac{z}{z+1} \right)^{\beta-1} \left(e^{-z^\beta} - e^{-(z+1)^\beta} \right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\liminf_{z \rightarrow \infty} \frac{\beta z^{\beta-1} \int_z^\infty e^{-t^\beta} dt}{e^{-z^\beta}} &\geq \liminf_{z \rightarrow \infty} \frac{\left(\frac{z}{z+1} \right)^{\beta-1} \left(e^{-z^\beta} - e^{-(z+1)^\beta} \right)}{e^{-z^\beta}} \\
&= \lim_{z \rightarrow \infty} \left(\frac{z}{z+1} \right)^{\beta-1} - \lim_{z \rightarrow \infty} \left(\frac{z}{z+1} \right)^{\beta-1} e^{z^\beta - (z+1)^\beta} \\
&= 1.
\end{aligned}$$

From Equation (24) we know that

$$\limsup_{z \rightarrow \infty} \frac{\beta z^{\beta-1} \int_z^\infty e^{-t^\beta} dt}{e^{-z^\beta}} \leq 1.$$

Hence, we conclude that

$$\lim_{z \rightarrow \infty} \frac{\beta z^{\beta-1} \int_z^\infty e^{-t^\beta} dt}{e^{-z^\beta}} = 1,$$

which implies that $m_{\delta, \beta}(z)$ is eventually non-positive for any $\delta \in (0, 1)$, i.e, assumption (e) holds for any $\delta \in (0, 1)$. □